

罗中泽

性别：男

年龄：23

邮箱：luozhongze0928@foxmail.com

电话：13607979355

个人主页：luozhongze.github.io



教育经历

香港中文大学（深圳） | 计算机与信息工程 | 哲学硕士（GPA: 3.85/4.00 排名: 4/54）

2025.09 - 2027.06

东北林业大学（211 工程） | 通信工程 | 工学学士（GPA: 88.08/100 排名: 14/73）

2021.09 - 2025.06

项目经历

2026.06-至今

面向行政执法全流程的可验证多智能体大模型系统

- 模型训练：基于 LLaMA-Factory 和 Qwen3.5-9B，设计四个任务专用 Agent，分别负责事实认定、法条匹配、处罚计算和决定说理；在昇腾 910C NPU 上冻结基座权重进行 SFT，仅训练注入注意力层、前馈网络层和投影层的 LoRA 参数。
- 数据治理与基准构建：每个任务构建均匀三领域的 30k 条训练样本，由真实数据和受控合成数据组成。合成阶段使用 Qwen3.5-27B 生成事实表述和决定理由，法条、处罚类型及金额均继承真实案件种子；另外冻结部分真实案件作为共同测试集，投影形成四个任务的 6k 条评测基准。
- 知识增强与多智能体管线：构建包含权威法规、裁量基准和真实历史案例的知识库，对 62k 条真实案例构建稀疏-稠密双路检索：使用 BM25 完成案由及法条的精确召回，使用 WeMM-Embedding-9B 向量模型检索语义相似案例，并通过 RRF 融合排序结果，将命中片段、来源案件及检索分数封装为可追溯证据包，为法条匹配、处罚计算及全局校验层提供可核验的知识依据；构建“事实认定→知识库证据检索→法条匹配→处罚计算→决定说理→文书模板生成与装配→全局校验”的 DAG。
- 可验证智能体与强化学习：构建全局校验 Verify 层校验跨智能体事实、法规时效、处罚结果及证据引用的一致性，采用 DPO 从候选偏好对中完成偏好对齐，引入 GRPO 强化学习后训练，奖励由本地正确性、格式可解析性与下游校验通过度加权构成，并设硬门禁控制高风险错误，提升系统合规性与稳定性。

科研经历

- Voices Across Registers: Corpus-Conditioned Vernacular Jailbreaks against Aligned LLMs via Fanfiction Subgenres. EMNLP 2026 Main. (第一作者, CCF-B)**

构建了一种基于语域的越狱框架，通过从 12 个真实同人小说子类型中选取段落作为创意写作提示的条件，并将有害请求嵌入叙事高潮中，无需针对目标的提示优化或攻击方大语言模型的生成。在 HarmBench 和 JailbreakBench 上使用四评委集成方法对 8 个对齐的大语言模型进行了大规模评估，并通过因子消融实验分离了语言语域相对于提示长度和结构的贡献。进一步提出了静态四轮扩展 VAR-A4，实现了 0.924 的平均攻击成功率（ASR），揭示了面向模板的安全防御的局限性。

- Multi-Physics: A Comprehensive Benchmark for Multimodal LLMs Reasoning on Chinese Multi-subject Physics Problems. ICASSP 2026. (第一作者, CCF-B)**

构建了 Multi-Physics，一个中文多模态物理推理基准，包含 1,412 道带图像的多项选择题，覆盖 11 个高中物理学科和 5 个难度等级。设计了结合最终答案准确性与针对思维链（CoT）逐步推理完整性的双重评估协议，并在纯文本和图文两种设置下对 20 个多模态大语言模型进行了基准测试。通过控制变量分析，考察了模型在学科、难度和输入模态上的表现差异，揭示了视觉定位与科学推理方面的不足，并公开了数据集和评估代码。

- KG2QA: Knowledge Graph-Enhanced Retrieval-Augmented Generation for Communication Standards Question Answering. ICASSP 2026. (第一作者, CCF-B)**

开发了 KG2QA，一个面向通信标准的端到端问答系统，基于 ITU-T 建议构建了 6,587 对问答数据集，并对 Qwen2.5-7B-Instruct 进行了微调。通过大模型辅助的三元组抽取（基于自定义本体）构建了包含 13,906 个实体和 13,524 条关系的领域知识图谱，并将其集成到知识图谱增强的检索增强生成（KG-RAG）流水线中，以实现知识检索和可溯源的答案生成。使用自动指标和 DeepSeek-V3 作为评判，与基础大语言模型及 RAG 基线进行了对比，最终通过网页平台和 API 将系统部署上线。

专业技能

- 大语言模型训练：SFT、LoRA/QLoRA、RL（PPO/DPO/GRPO 等）；vLLM、LLaMA-Factory、verl。
- 检索增强生成与知识图谱：Neo4j、Cypher、本体设计、三元组抽取、Ollama、LangChain、向量检索、BM25。
- 评估：测试基准构建、LLM-as-a-Judge、CoT 评估；指标包括 ASR、F1、BLEU、ROUGE。
- 工程能力：Python、Linux、PyTorch、Transformer、FastAPI、Git、Vibe Coding 经验丰富。
- 语言能力：IELTS 6.5、英文论文撰写与 LaTeX 排版经验丰富。